README
Partea 2 - PCLP3 - Titanic

Cerinta 1

Definim o functie care aplica IQR pe o coloana dintr-un set de date, conform cerinta. Un argument important al functiei este negativeData:

- daca este 0 => nu are sens ca intervalul returnat de IQR sa aiba un capat negativ (exemplu: varsta)
- daca este 1 => are sens ca intervalul returnat de IQR sa aiba un capat negativ

Un alt argument este col care contine numele coloanei pe care aplicam IQR

Cerinta 2

Definim o functie care aplica Z-Score pe o coloana dintr-un set de date. Functia care calculeaza Z-Score este predefinita(precum si cea pentru IQR). Argumentele acestei functiei, pe langa setul de date, sunt numele coloanei pe care se aplica Z-Score si toleranta(valoarea de la care informatia este considerata un outlier).

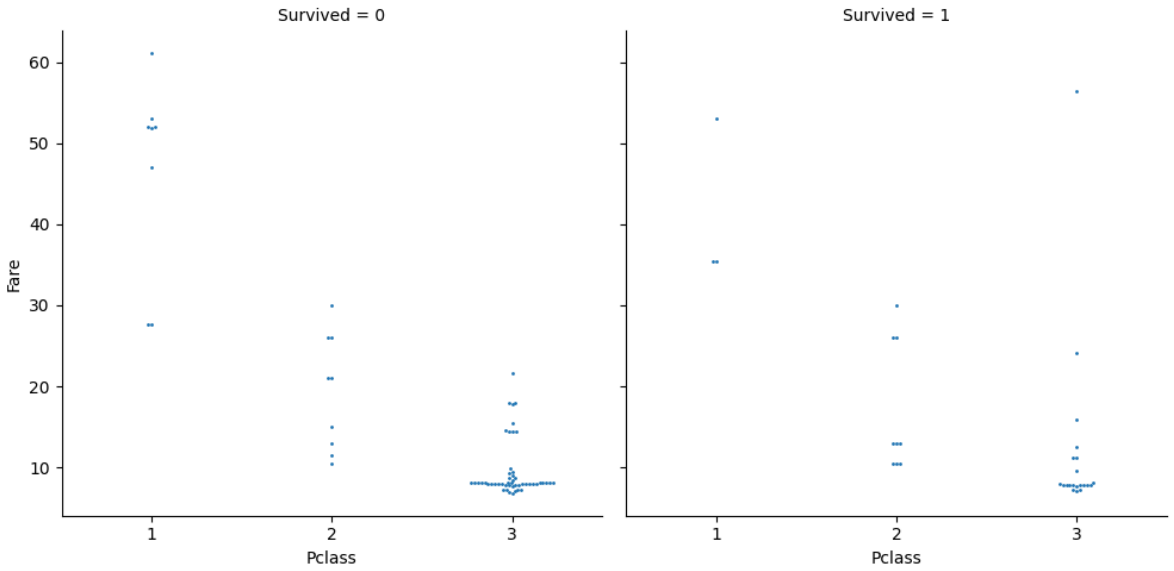
Cerinta 3

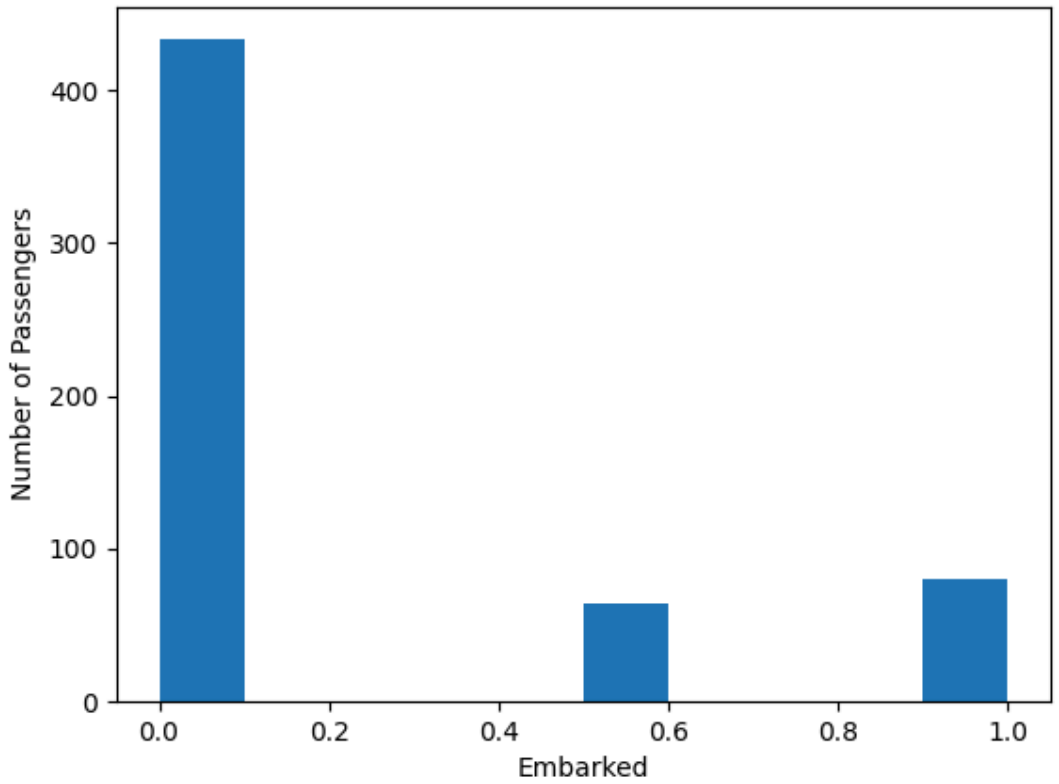
In procesul de eliminare al outlierilor am aplicat:

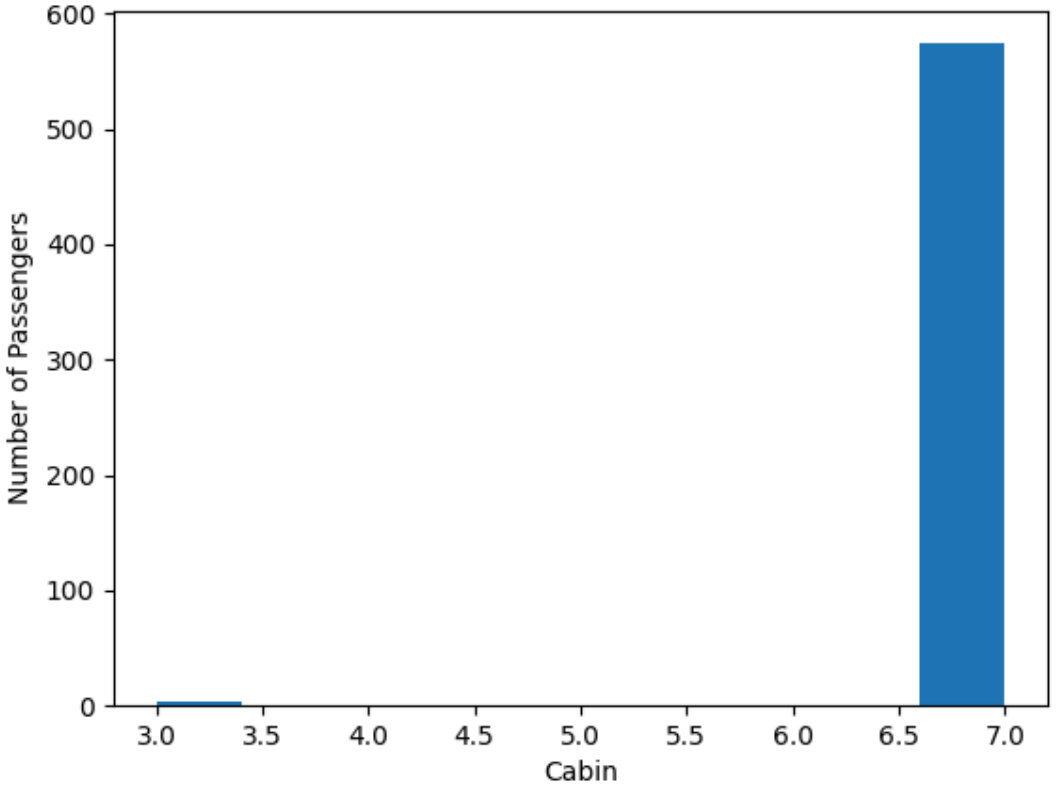
- IQR pe coloana ‘Ages’, pentru a elimina valorile extreme
- IQR pe coloana ‘Fare’
- Z-Score pe coloana ‘SibSp’
- IQR pe coloana ‘Parch’

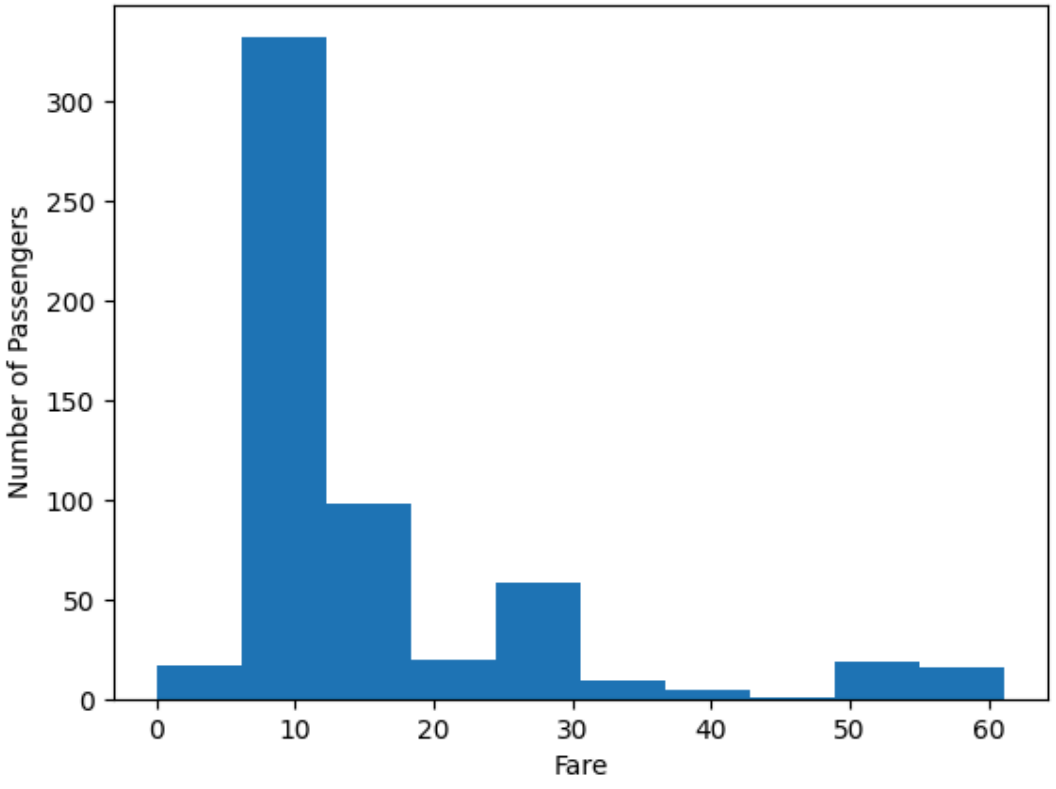
Incarcam fisierul de tip csv, salvat in urma eliminarii outlierilor. Aplicam functiile definite la partea I si comparam graficele obtinute. In urma comparatiei, se constata cu usurinta ca distributia datelor nu a most modificata semnificativ, observandu-se doar lipsa outlierilor.

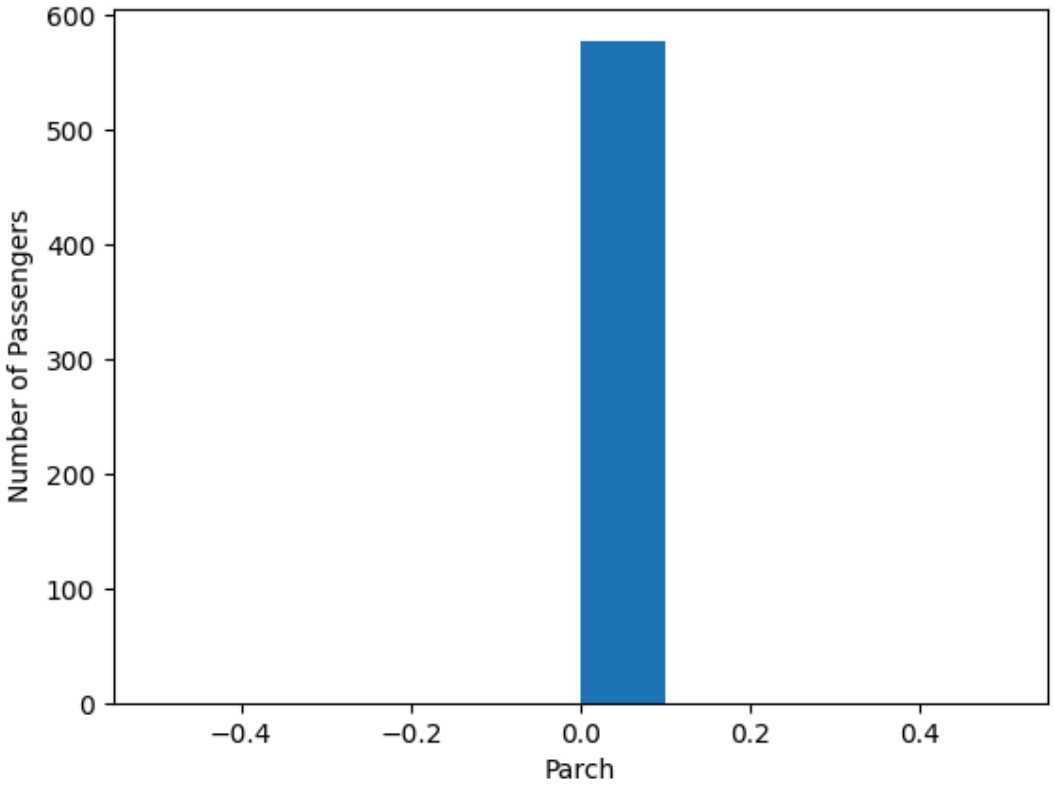
In urma eliminarii datelor s-au generat urmatoarele grafice:

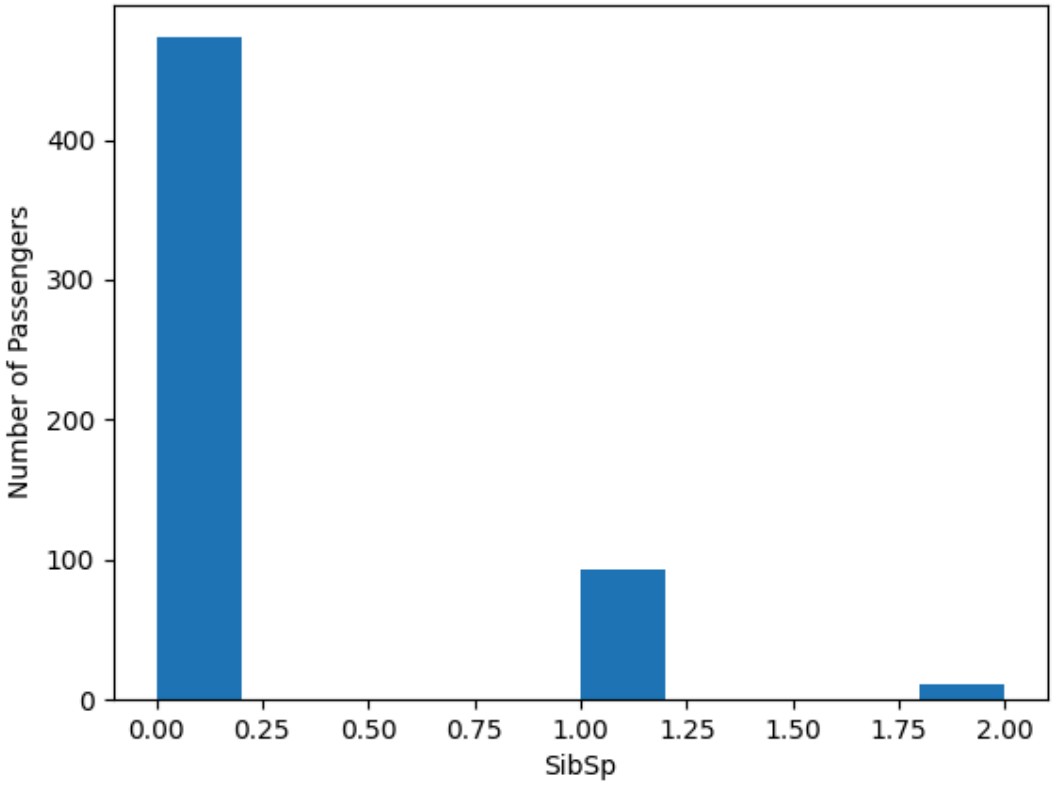

Survived = 0
Survived = 1
Fare
60
50
40
30
20
10
1
2
3
Pclass

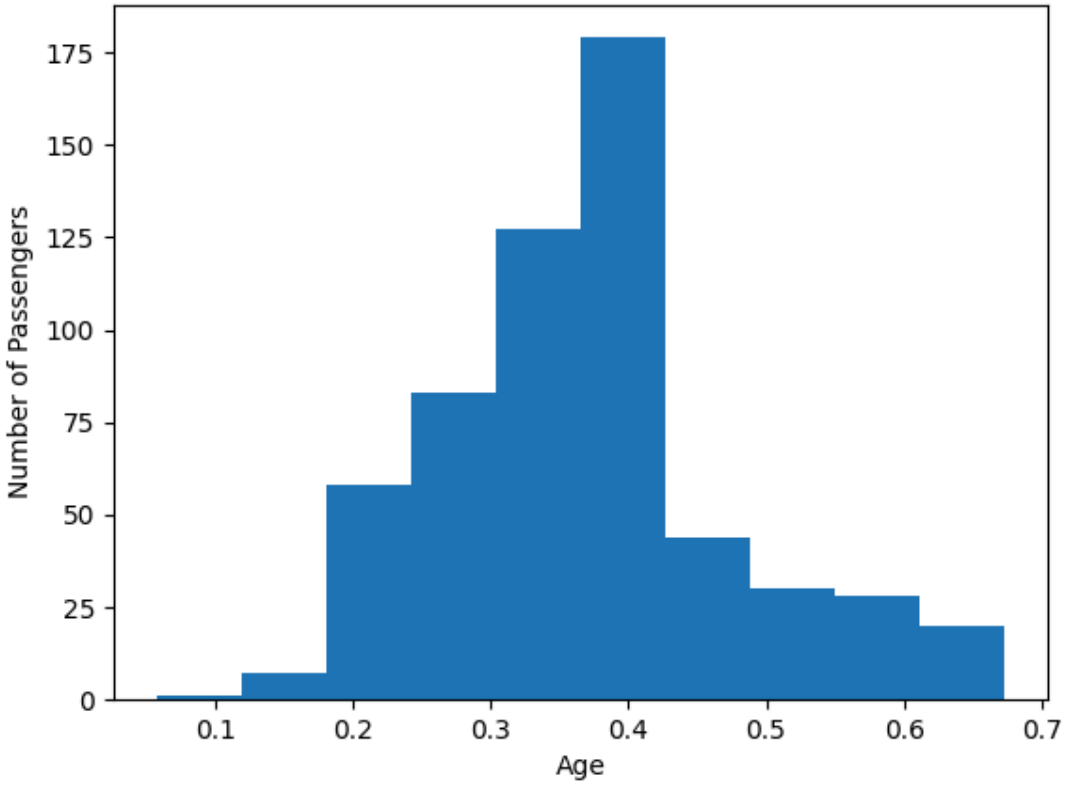

Number of Passengers
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
Embarked

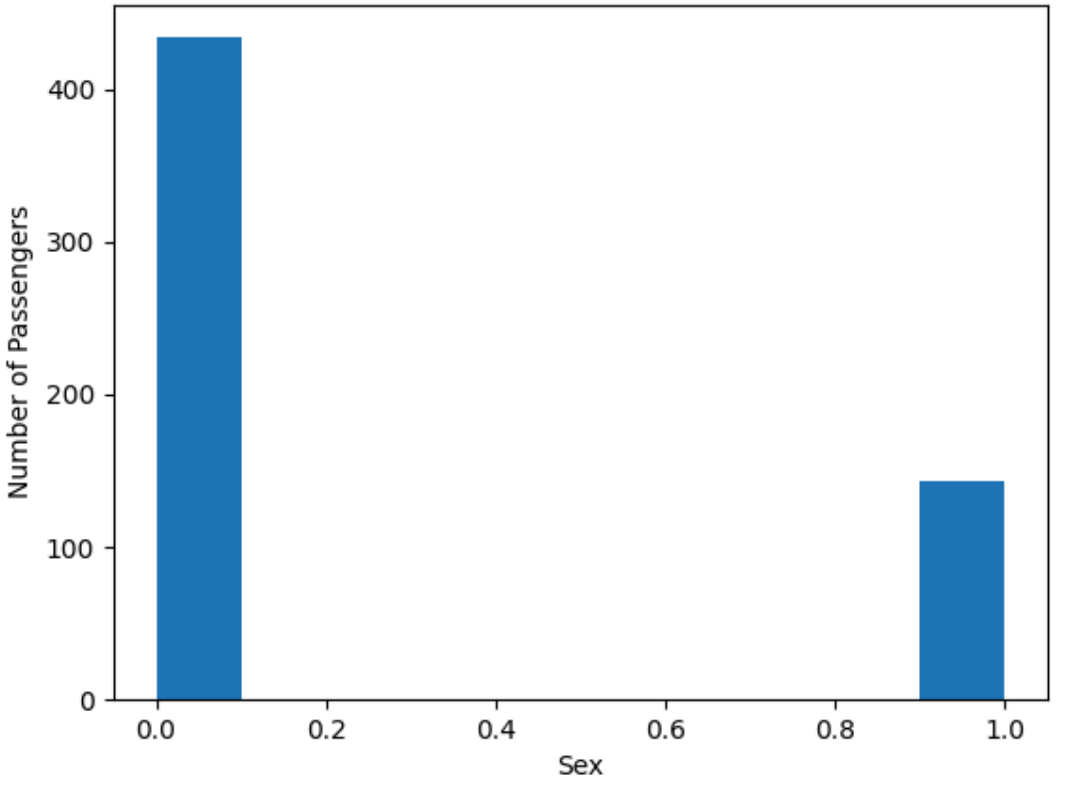

600
500
400
300
200
100
0
Number of Passengers
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Cabin

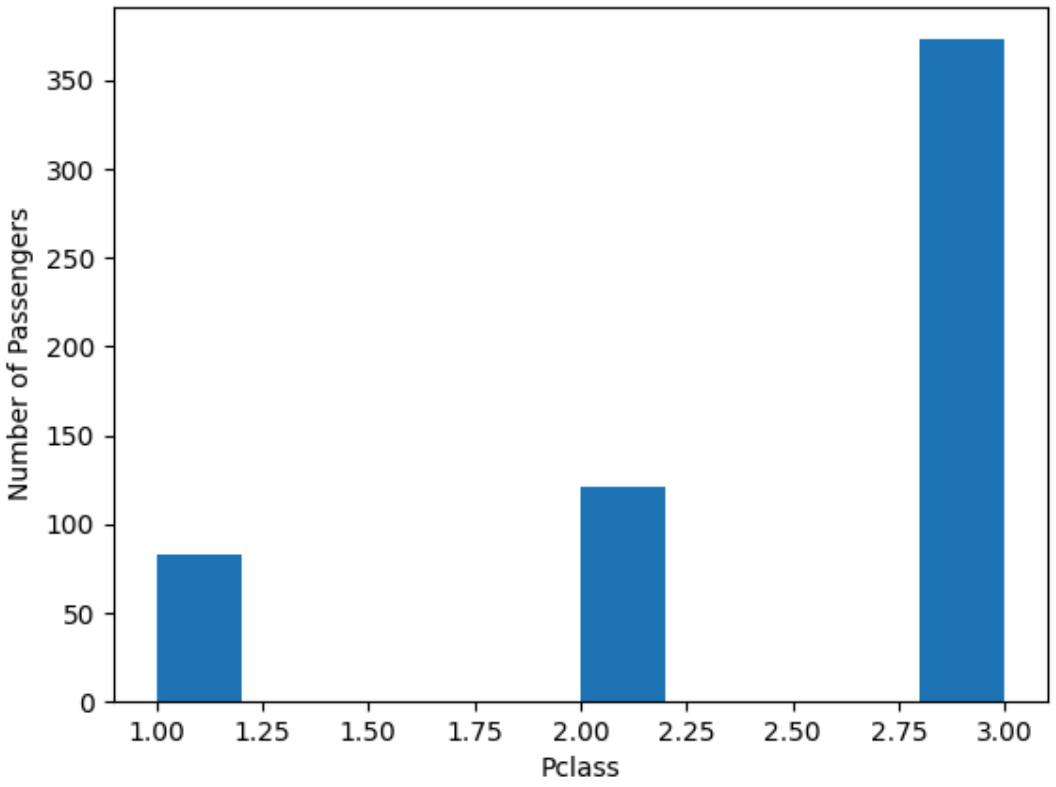

300
250
200
150
100
50
0
Number of Passengers
0
10
20
30
40
50
60
Fare

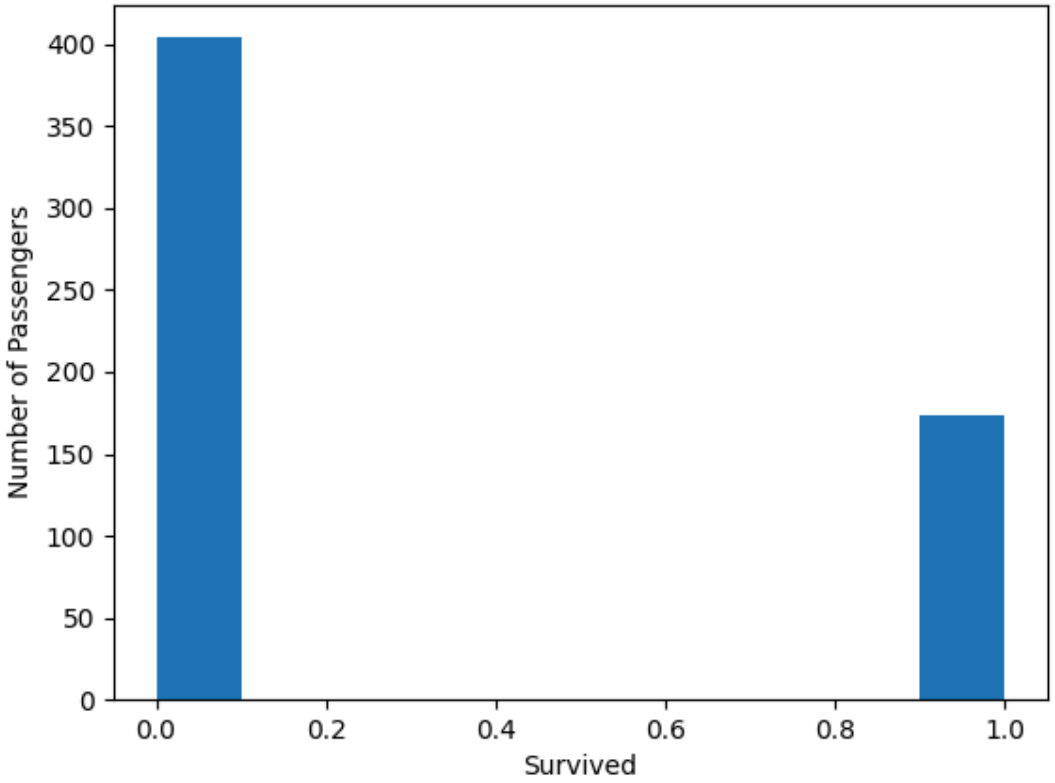

600
500
400
300
200
100
0
Number of Passengers
−0.4
−0.2
0.0
0.2
0.4
Parch

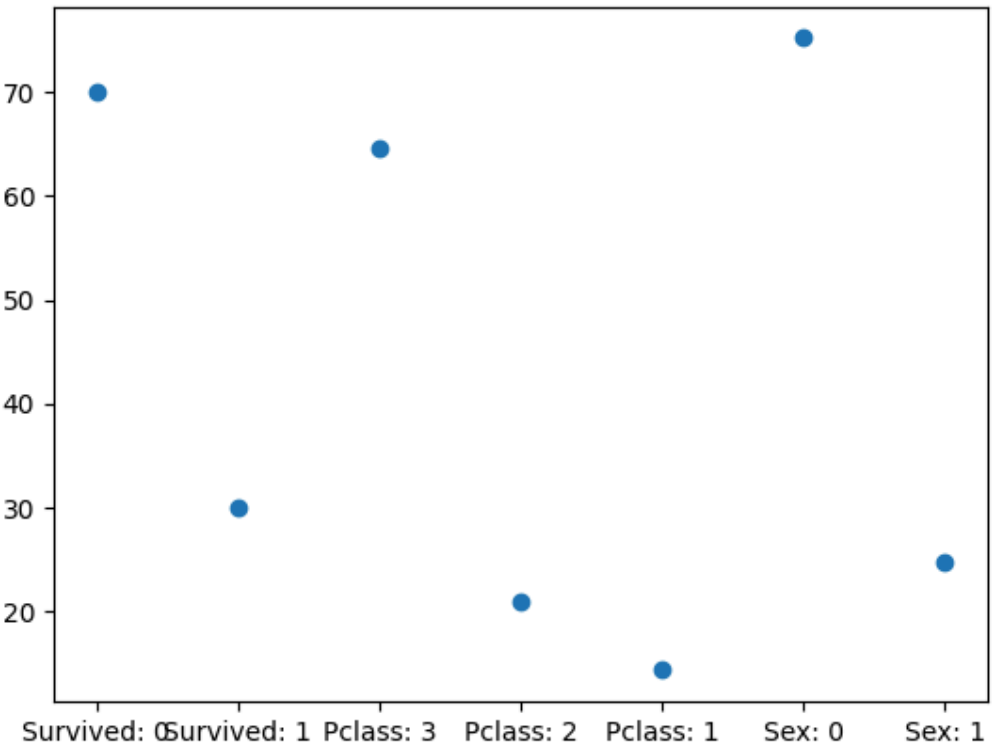

400
300
200
100
0
Number of Passengers
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
SibSp


175
150
125
100
75
50
25
0
Number of Passengers
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Age


400
300
200
100
0
Number of Passengers
0.0
0.2
0.4
0.6
0.8
1.0
Sex


Number of Passengers
0
50
100
150
200
250
300
350
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Pclass

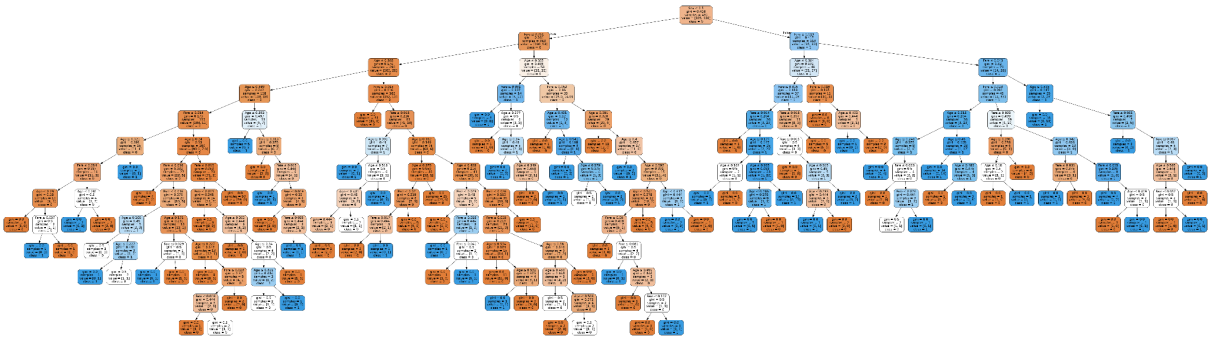

Number of Passengers
400
350
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
Survived


70
60
50
40
30
20
Survived: 0
Survived: 1
Pclass: 3
Pclass: 2
Pclass: 1
Sex: 0
Sex: 1

Cerinta 4

Pentru realizarea cerintei 4, am urmarit pasii din enunt. Pentru a evalua acuratetea modelul generat am folosit 20% din setul de train si setul de test, deoarece din setul de train am eliminat outlierii, iar in setul de test acestia exista. Astfel, se obtin valori diferite. Pentru generarea modelului am folosit un Decision-Tree generat de functii din sklearn. Pentru o interactiune mai usoara cu interfata modelul, am facut o interfata grafica din care putem alegem ce coloane sa fie folosite de model si ce date din ce coloane sa fie normalizate.

Pentru un exemplu de rulare cu ‘Age’, ‘Fare’, ‘Parch’, ‘Cabin’, ‘Sex’ selectate si ‘Age’ si ‘Fare’ normalizate, arborele de decizie este:



Pentru acelasi exemplu, se obtin urmatoarele date:

Accuracy on 20% of training data no outliers: 0.8362068965517241 5.903701848217464

Accuracy on test data: 0.7942583732057417 7.415679883885347